DOCUMENT RESUME

ED 170 045 **PS 010 545**

AUTHOR Lichtenstein, Robert

TITLE Classificational Methods for the Psychometric Evaluation of Screening Procedures.

PUB DATE Mar 79
NOTE 13p.; Paper presented at the Biennial Meeting of the Society for Research in Child Development (San Francisco, California, March 15-18, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Decision Making; *Identification Tests; Psychometrics; Reliability; Research; *Screening Tests; *Statistical Analysis; Statistical Data; Validity
IDENTIFIERS *Qualitative Data

ABSTRACT

The trend towards using screening programs to facilitate early identification of developmental or learning problems has accentuated the need to establish reliable and valid screening procedures. Researchers in this area must recognize that the dichotomous decision-making function of screening (i.e., to refer or not refer a child for further assessment) requires psychometric evaluation methods which are suited to this function. Traditional reliability and validity coefficients, frequently applied to screening procedures, are designed to express relationships between continuously-scored predictors and criteria. For a screening procedure, however, only classificational (i.e., group membership) distinctions are of consequence; measurement variance within the group of referred or non-referred individuals is irrelevant in evaluating a discrete screening outcome. A contingency table showing the cross-tabulation of a dichotomous screening predictor and a dichotomous criterion variable is commonly used to represent screening decision hit rate. A standard index for summarizing the predictor-criterion relationship in such a table has yet to be established. In reviewing alternatives, a measure which strategically weights the false positive and false negative error rates seems preferable. A cross-classificational model for dichotomous test and criterion measure and a compendium of cross-classificational indexes are included. (Author/RH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED170045

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CLASSIFICATIONAL METHODS FOR THE PSYCHOMETRIC EVALUATION OF SCREENING PROCEDURES

Robert Lichtenstein

University of Minnesota

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert Lichtenstein

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

PS 010545

Paper presented at the biennial meeting of the Society for Research in Child Development, San Francisco, March 15, 1979.

Classificational Methods for the Psychometric

Evaluation of Screening Procedures

Despite the obscure title, this paper addresses a measurement topic which is pertinent to a wide range of clinical, educational, and research applications--the appropriate presentation of validity and reliability data for tests whose use is based on classificational scores. By classificational, I am referring in this paper to categorical or qualitative scores (e.g., the set of scores below average, average, and above average) as opposed to continuous raw scores or standardized scores. To introduce the topic, let me present my conclusions first:

1. First, it is important to recognize the type of information upon which an index of reliability or validity is based, and the interpretations that may follow from its use.

2. Second, validity coefficients (i.e. Pearson correlations) don't tell you what you need to know about a test's value for making discrete decisions. Screening measures provide a rather clear-cut example of tests which call for classificational methods of evaluation.

3. Third, even if we come to some agreement about using a classificational model where appropriate, a standard manner of statistically summarizing such information has yet to be resolved.

4. Finally, the methodological and theoretical sources that provide the basis for this paper have been around for at least 20 years, so I won't get too discouraged if the perfectly reasonable

recommendations that follow continue to be disregarded.

With recent federal and state mandates requiring school districts to provide early intervention services for preschool children with special educational needs, there has been a proliferation of screening programs to facilitate early identification efforts, and a corresponding proliferation of screening tests intended for use in these screening programs. Unfortunately, screening program administrators responsible for selecting developmental or educational screening tests find that the scarce psychometric data that do exist are presented in a manner that gives little indication of how accurate and efficient the tests are for their intended purpose. The traditional approach to psychometric evaluation, which uses correlations to express the linear relationship between continuously distributed predictor and criterion variables, may be appropriate for tests where the exact placement of a score within a distribution is of interest. However, a screening test is not intended for diagnostic use; its sole function is to indicate whether an individual is or is not in a high risk classification and warrants further assessment. Measurement variance within a classificational group, e.g., the difference between a score just above the cutoff and a very high score, contributes to a correlation coefficient, but is simply irrelevant to the dichotomous decision to refer or not to refer a child for assessment. Applying a classificational model of psychometric evaluation, the reliability of a screening test is a function of how consistently a screening test will generate the same classificational outcome. The validity of a screening test is measured by comparing the decision of whether or not

to refer an individual with the person's actual status, i.e., their need for treatment as determined by further evaluation or by later manifestation of the problem that was screened for.

Table 1 of the handout illustrates the cross-tabulation of a dichotomous predictor variable (screening test outcome of either high risk or low risk) and a dichotomous criterion variable (intervention needs or no intervention needs) in a fourfold contingency table. The cells in the upper left and lower right represent hits, or agreement between the screening test and the criterion measure. The lower left cell represents false negatives, or under-referrals, i.e., the failure of the screening test to refer a person who is in need of intervention, and the upper right cell represents false positive errors, or over-referrals to assessment.

On the other side of the handout is an example comparing a correlational summary (Table 3) and a classificational summary (Table 4) of the results of a concurrent validity study. In this study, two screening tests and three assessment measures were administered to 40 Minneapolis preschool children as part of the validation research for the Woodcock-Johnson Psycho-Educational Battery. Here, I present only the data relating Stanford-Binet results with the two screening tests--Developmental Indicators for the Assessment of Learning (DIAL) and the Denver Developmental Screening Test (DDST).

In comparing the two screening tests to the Stanford-Binet, the validity coefficients were quite favorable. These correlations were derived from continuously-scored variables--DIAL total scaled score, DDST total raw score, and Stanford-Binet mental age. The DDST and

Stanford-Binet correlated .65, and the DIAL and Stanford-Binet correlated .74. Both of the screening tests, however, use classificational scores for decision-making purposes which are derived through a non-linear combining procedure, in other words, a decision rule which is more involved than just dichotomizing the total score distribution with a cutoff point. The two diagrams on the left in Table 4 show the agreement between dichotomized Stanford-Binet scores with a cutoff point at IQ of 80, and classificational scores for each screening test. Two children fell into the "intervention needs" category with Stanford-Binet scores below 80. Reading across the top line, the DDST referred four children, only one of which was a hit, or agreement with Stanford-Binet classification. In the table below, the DIAL referred only a single child, but that was a hit. Predictive efficiency of the screening tests is less impressive when presented in this manner. The most surprising result, however, comes from the comparison of DIAL and DDST outcomes shown in the table on the right. Although the continuous scores have a strong linear relationship, with a whopping correlation of .82, there was absolutely no agreement between the referrals made by the two different screening tests. Clearly, it makes a big difference whether the DIAL or the DDST is used for screening at this particular preschool center. Nevertheless, the validity data supplied for these two screening tests are reported in terms of validity coefficients and based on the continuous scores. This, sad to say, is the rule rather than the exception in validating screening tests.

To document this statement, a total of 115 developmental screening tests for preschool aged children were reviewed, all of which have

been in distribution since 1976. Test-retest or inter-rater reliability data were available for 51 tests, of which only one presented evidence of reliability of the classification score. Validity data were found for 36 of the screening tests in either the test manual or in the literature, of which only two studies were presented in terms of classificational agreement.

Screening procedures, with their uniquely well-defined function, make an easy target for such criticism, but it is common for various other measures which are applied towards decision problems to be inappropriately evaluated. An institution which uses a particular measure for making classificational decisions, such as an IQ test to determine special class placement, needs information about the decision-making efficiency of the test in order to evaluate and optimize the benefit derived from a particular decision rule. Research measures may also involve classification, as in assigning individuals to different levels of an experiment. For example, a measure of depression may be dichotomized with a median split to established a depressed and a non-depressed group. The reliability of such a measure should not reflect measurement error averaged over the entire continuum, as would be indicated by a correlation coefficient, but rather the discriminating power of the measure at the cutoff point.

In stressing the importance of classificational methods for evaluating screening tests and other decision-making procedures, I haven't been very specific about how to analyze a cross-classificational table in order to compare alternative tests or to select optimal cutting points. Basically, there is no standard index for summarizing

the degree of association between two variables in a 2 x 2 contingency table. Without venturing off into the statistical hinterlands, let me present a quick overview of what I call the "search for the ultimate cross-classificational index." Some of the more interesting or widely used entries in this competition are described below Table 1 with cookbook formulas, and these statistics are computed for the sample data in Table 2 (adapted from Feshbach, Adelman, and Fuller, 1977).

Total hit rate, the most straightforward manner of expressing agreements between predictor and criterion, has the distinct disadvantage of being dependent upon category base rates, which gives misleading results when base rates are extremely uneven (Meehl & Rosen, 1954). In the example, total hit rate is 90%, but a decision rule that predicts adequate reading performance for everyone hits 91% without the test.

In order to express hits relative to base rates, a pair of statistics can be used to indicate the proportion of each type of hit, true positives and true negatives, relative to category base rates. Co-positivity and co-negativity, also known as sensitivity and specificity when the criterion measure is a "true diagnosis" rather than a fallible reference test, express the proportion of true positives and true negatives relative to criterion measure base rates (Buck & Gart, 1966). In a similar manner, Bayes' Theorem provides the basis for a pair of predictive utility indexes which indicate the likelihood of the expected outcomes corresponding to each prediction contingency, which translate as the proportion of true positives and true negatives relative to pre-

diction measure base rates. Predictive utility of a positive screening result is analogous to co-positivity as an index of the false positive rate, and predictive utility of a negative screening result is comparable to co-negativity. The figures in each pair are interdependent--changing a cutoff score to reduce false positives lets in more false negatives, and vice versa.

A single index is desirable when you need to select one decision procedure over another. Numerous statistics have been proposed which manipulate the cell frequencies (a, b, c, and d in Table 1) to generate an index which ranges from 0 to 1, or -1 to 1, but these tend to lack interpretability and weight false positive and false negatives errors equally. Perhaps the most interpretable single index is phi coefficient, the Pearson correlation for the special case of two dichotomous variables. Proposed for use in comparing alternative cutoff points, phi coefficient is greatest at the cutting point which maximizes the sum of co-positivity and co-negativity (Guttman & Raju, 1965). The index of performance (Youden, 1950) is derived from and directly related to the simple sum of co-positivity and co-negativity. The coefficient of association is one of several indexes developed by Yule (1912) for their mathematical properties, and also weights false positive and false negative errors equally.

In an actual decision-making process, each type of error may be associated with different costs and benefits. For example, in screening, a false positive error can be corrected following referral for further assessment, whereas false negatives constitute a failure to receive treatment and are usually of greater consequence. Cronbach and Gleser

(1957) propose a payoff matrix to evaluate a decision rule, in which each predictor-criterion contingency, or cell, in our model is weighted by the relative costs and benefits associated with that course of action. This places upon test users the responsibility of assigning price tags to difficult-to-characterize circumstances, such as the benefit derived from placing a developmentally disabled child in a preschool intervention program, or the cost to a child who is inappropriately labeled and assigned to that same program. Still, it seems feasible to estimate the relative importance of false negatives and false positives in a particular situation, and this can provide the basis for computing a weighted combination of co-positivity and co-negativity or of the two predictive utility indexes as a measure of decision-making efficiency. Such an index might differ with each particular test use, but that still makes perfect sense to me. That's my leading choice to date, but I'd be delighted to get some other opinions on the matter.

The rationale for classificational methods to evaluate screening tests and other measures which use classificational scores seems so self-evident that I am almost tempted to apologize for being trivial. However, inappropriate and uncritical use of correlation coefficients is so prevalent that a reminder of this sort is necessary every 20 years or so.

References

Buck, A., & Gart, J. Comparison of a screening test and a reference test in epidemiological studies. American Journal of Epidemiology, 1966, 83, 586-592.

Cronbach, L., & Gleser, G. Psychological tests and personnel decisions. Urbana, Ill.: University of Illinois Press, 1957.

Feshbach, S., Adelman, H., & Fuller, W. Prediction of reading and related academic problems. Journal of Educational Psychology, 1977, 69, 299-308.

Guttman, I., & Raju, N. A minimum loss function as determiner of optimal cutting scores. Personnel Psychology, 1965, 18, 179-185.

Meehl, P., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or acting scores. Psychological Bulletin, 1955, 52, 194-216.

Youden, W. Index for rating diagnostic tests. Cancer, 1950, 3, 32-35.

Yule, G. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society, 1912, 75, 579-642.

Table 1. Cross-classificational model for dichotomous screening test and criterion measure

		Criterion Measure: Intervention Needs (IN)	Criterion Measure: No Intervention Needs (NIN)	
Screening Test	High Risk +	True positives a	False positives c	Total Positives a + c
	Low Risk -	False negatives b	True negatives d	Total Negatives b + d
		Total IN a + b	Total NIN c + d	Total a + b + c + d

Table 2. Sample data (from Feshbach, Adelman, and Fuller, 1977)

	Reading Level: Low	OK	Total
+	19	20	39
-	25	387	412
T	44	407	451

A compendium of cross-classificational indexes

Total hit rate:	$H_T = \frac{a + d}{a + b + c + d}$	$H_T = .90$
Co-positivity:	$C_1 = \frac{a}{a + b}$	$C_1 = .43$
Co-negativity:	$C_2 = \frac{d}{c + d}$	$C_2 = .95$
Predictive utility, screening positive:	$p\,(IN +) = \frac{a}{a + c}$	$p\,(+) = .49$
Predictive utility, screening negative:	$p\,(NIN -) = \frac{d}{b + d}$	$p\,(-) = .94$
Phi coefficient:	$\phi = \frac{ad - bc}{(a + b)(c + d)(a + c)(b + d)}$	$\phi = .40$
Index of performance:	$J = \frac{ad - bc}{(a + b)(c + d)}$	$J = .38$
Coefficient of association:	$\lambda = \frac{ad - bc}{ad + bc}$	$\lambda = .87$

Table 3. Correlational model: concurrent validity of two preschool developmental screening tests

	DDST	DIAL	Stanford-Binet
DDST	---		
DIAL	.82	---	
S-B	.65	.74	---

Table 4. Classificational model: concurrent validity. . . of two preschool developmental screening tests

		Stanford-Binet		
		Below 80	80+	Tot.
DDST	+	1	3	4
	-	1	34	35
	Tot.	2	37	39

		DIAL		
		+	-	Tot.
DDST	+	0	4	4
	-	1	35	36
	Tot.	1	39	40

		Stanford-Binet		
		Below 80	80+	Tot.
DIAL	+	1	0	1
	-	1	37	38
	Tot.	2	37	39